PravdaPulse

A Database for Sentiment Analysis of Russian News Articles since 2021

Codebook

Tessa Conrardy

IGA 677: Russian National Security Policy
Visiting Associate Professor Yuri M. Zhukov
December 11, 2023

Introduction

While the notion of "national security" usually elicits imagery of fortifications and armaments, 21st-century warfare is also inevitably shaped by mass media and the ways that we consume it. Control over the narrative of a conflict can prove just as important as control over the lines on the battlefield.[1] With the inflation of its propaganda machine and the eradication of broad swaths of Russia's independent media, the Putin regime has shown just how significant media freedoms—or lack thereof—can be in a time of war.[2]

While the escalating persecution of Russia's independent press since the annexation of Crimea (and especially since the full-scale invasion of Ukraine) has raised alarm bells in the international community, the scholarly community has struggled to keep pace in analyzing changes in the content and tone of Russian news media in the face of these broader structural changes.[3] To this end, some helpful media-parsing projects—notably, the "Media Cloud" project right here at Harvard—have emerged to help facilitate large-n analysis of news media.[4] These tools, however, are often limited to cursory textual analysis tools like word frequency; only with

---

[1] Eschenauer-Engler, Tanja. "Armed Forces and Airwaves: Media Control and Military Coups in Autocracies." *Contemporary Politics* 29, no. 4 (August 8, 2023): 446–65. https://doi.org/10.1080/13569775.2023.2173874;
Hill, Raymond R., Jr. "The Future Military-Media Relationship: The Media as an Actor in War Execution." Air Command and Staff College, March 1, 1997. https://apps.dtic.mil/sti/citations/ADA394012;
Norris, P., and R. Inglehart. "Silencing Dissent: The Impact of Restrictive Media Environments on Regime Support," 2008. https://www.semanticscholar.org/paper/Silencing-dissent-The-impact-of-restrictive-media-Norris-Inglehart/50e2915b34e9a1f3c144068d49bf736ae58dd677;
Payne, Kenneth. "The Media as an Instrument of War." *The US Army War College Quarterly: Parameters* 35, no. 1 (March 1, 2005). https://doi.org/10.55540/0031-1723.2243.

[2] Giles, Keir. "Russia's Hybrid Warfare: A Success in Propaganda." Federal Academy for Security Policy, 2015. https://www.jstor.org/stable/resrep22215,
Mylchenko, Larysa. "The Russian Influence through Mass Media as a Significant Factor of Hybrid War against Ukraine." *Вісник Книжкової Палати*, no. 10 (October 28, 2021): 8–16. https://doi.org/10.36273/2076-9555.2021.10(303).8-16,
Klarić, Darijo, and Josip Mandić. "Case Study of the Russian Disinformation Campaign During the War in Ukraine – Propaganda Narratives, Goals and Impacts." *National Security and the Future* 24, no. 2 (July 4, 2023): 97–139. https://doi.org/10.37458/nstf.24.2.5.

[3] Paul, Christopher, and Miriam Matthews. *The Russian "Firehose of Falsehood" Propaganda Model: Why It Might Work and Options to Counter It*. RAND Corporation, 2016. https://doi.org/10.7249/PE198.

[4] Berkman Klein Center for Internet and Society, Harvard University. "Media Cloud," n.d. https://www.mediacloud.org/.

access to full article texts can researchers pursue more nuanced methods like sentiment analysis. A divide surely exists between the coverage of state-coopted news outlets and exiled independent counterparts, but how does that divide manifest? What themes are these different outlets focusing on in their coverage? How much variation exists in the tone or focus of coverage among state-run outlets? Among independent outlets? The PravdaPulse database is designed to help answer questions like these by providing large volumes of article text from several Russian news outlets in a format suitable for sentiment analysis.

While this database is inherently limited and must (for reasons outlined in the following pages) be taken with several grains of salt, my hope is that this data can serve as an early foundation for more careful, empirical examination of the ways that the Russian media ecosystem is evolving in an increasingly convoluted era of information politics.

Methodology

Compilation of this database predominantly relied on two tools: the Chrome-based web-scraping platform "Web Scraper" and the statistics platform RStudio. Using these tools, I executed the following steps:

1. **Setup.** I started by choosing a selection of Russian news outlets to focus on—namely Dozhd, Channel 1, Kommersant, and Pravda.[5] Four outlets may sound conservative, but in documenting every article that these outlets have published between January 2021 and November of 2023, the PravdaPulse database stacks up to hundreds of thousands of articles.[6]

2. **Scraping.** Once I had decided on a list of outlets, I then configured Web Scraper so that when provided with a preliminary URL for a given outlet, Web Scraper could comb through an outlet's archive and collect data for every article through a time range of my

[5] With that said, the data collection process described in this codebook could be replicated by other researchers for additional outlets of interest.
[6] Note that scraping collected all articles between January 1, 2021, and November 24, 2023.

choosing. (Scraped data included each article's headline, date, article text, and, when provided, content section and number of views.) This step entails significant time spent on configuration and processing for each outlet.[7,8,9]

3. **Cleaning.** Once data had been scraped, I used RStudio to perform basic data cleaning functions like standardizing date formats, removing unnecessary formatting holdovers (e.g. the erroneous paragraph-separating characters that Web Scraper adds between <p> tags).[10]

4. **Uploading Datasets via Github.** While I had originally planned to combine data from all outlets into one dataset, it quickly became clear that the sheer volume of data demanded multiple files. Given that even each outlet had multiple megabytes worth of data, I chose to divide data into csv files by outlet and year. (E.g. all of the Channel 1 articles from 2021 can be found in the "Channel1_2021.csv" file.) Both for ease of processing and because of Github's 25 MB file size restriction, any outlet with over 5,000 articles per year was broken down into alphabetized 5,000-entry datasets. (E.g. Kommersant's 2023 articles are broken down into "Kommersant_2023_*A*.csv" through "Kommersant_2023_*I*.csv".) These datasets are all available for download via the PravdaPulse Github repository. This arrangement also allows users to download portions of the data that are more relevant to their work, (e.g. only 2022 data or only Pravda data) saving space and processing power.

[7] On the note of configuration, each site has slightly a different layout, and Web Scraper must be trained to the distinct structure of each site. Familiarity with html is helpful to this end, but Web Scraper also has a more user-friendly selection tool detailed in the platform's documentation:
Web Scraper. "Documentation - Selectors," 2023. https://www.webscraper.io/documentation/selectors.
[8] Depending on the structure of the site, I predominantly scraped either by section (e.g. "Экономика," "Мир," "Спорт," etc.) or by year to split processing into more manageable "chunks." See further documentation in Appendix A.
[9] Web Scraper runs at a pace of about two seconds per article scraped—hardly breakneck speed, but much faster and more convenient than compiling information for thousands of articles manually. Other platforms are capable of scraping more quickly, but Web Scraper is specifically designed to prevent site overloading. Sending too many requests to a site per second is actually a strategy often used to intentionally crash sites, (this kind of site-crashing is otherwise known as a Denial of Service, or DoS, attack) so Web Scraper, while more time-intensive, avoids this potentially harmful strain on scraped sites by putting a buffer period between each request.
[10] Find the relevant R code on the Github repository.

Variables

See below a list of each variable collected for a given article. I also provide some usage notes (i.e. formatting nuances, sampling considerations, etc.) and the platform used to collect each variable.

`Outlet_Name`	English-language iteration of news outlet name. *If outlet did not provide its own iteration of an English-language name, a transliterated version of the Russian-language name was provided. (This variable subsequently contains no cyrillic characters.)* *Collected manually from the header or footer of each outlet's home page.*
The outlets detailed in the database are Channel 1, Dozhd, Kommersant, and Pravda.	
`Section`	The section under which the article is categorized. *Some of the outlets that I examined do not have sections or do not provide the section under which an article was originally published. Others sometimes provide multiple relevant sections. (E.g. "Culture" and "Politics.") Recorded as a comma-separated list in cases of multiple sections. Recorded as an NA value when no section was provided.* *Collected via the Web Scraper platform, sometimes by parsing for the relevant html class and sometimes by pulling from the structure of the sitemap. (In cases where archives are organized by section—namely Kommersant.)*
`Article_Link.href`	An article-specific URL. *Collected via the Web Scraper platform.*

`Headline`	Headline of the article. *Note that in some cases, outlets provided secondary headlines. When possible, I opted to incorporate these into the* `Article_Text` *variable, but some niche instances of these secondary headlines may have been excluded from both the* `Headline` *and* `Article_Text` *variables, constituting a potential minor blind spot in this methodology.* *Collected via the Web Scraper platform, usually by parsing for the relevant html class. (E.g. <H1>, <h1.title>, etc. depending on site-specific convention for headlines.)*
`Article_Text`	The full text of the article. *Note that for layouts that diverge from the standard article layout, unusual elements like pull quotes or embedded social media posts may not be captured as part of the article text, though I incorporated these conventions when I could identify them.* `Article_Text` *may also, on the other hand, include minor extraneous elements like a reporter sign-off at the end of the article. These minor instances of extraneous text may add some (likely negligible) additional noise to textual analysis.* *Also note that some unconventional formats like interviews could potentially skew data in unpredictable ways. (E.g. many occurrences of an interviewee name because the name is stated before every one of the interviewee's interview responses.)* *Collected via the Web Scraper platform, usually by parsing for the relevant html class. (Usually some variation of <p>, with some exceptions.)*
`Date`	Date when the article was published. *For some platforms, dates are updated per the most recent instance when an article was updated, as opposed to when the article was originally*

	uploaded. This is a minor caveat that should be taken into consideration when incorporating the "Date" variable into analysis. *Collected via the Web Scraper platform, usually by parsing for the relevant html class. (E.g. div.date, etc. depending on site-specific convention for publication date.)*
`Views`	The number of views for a given article. *This metric is not always provided, but is collected for sites that do provide it and listed as an NA value for sites that do not provide it.* *Collected via the Web Scraper platform, usually by parsing for the relevant html class. (E.g. .iconbox__item--viewing or equivalent.)*

Limitations and Further Research

Though the database described in this codebook contains a large volume of data well-suited for sentiment analysis regarding Russian news trends, there are several fronts on which a researcher using this database should take pause. For one, this database is not necessarily representative of the broader Russian media ecosystem. Though I draw on several influential outlets, there are many, many other outlets publishing news and current events content for the Russian audience. Thus, while these data may help illuminate some potential trends, this database should not necessarily be assumed to be a representative sample of circulating mass media, nevermind interpreted as an exhaustive or definitive survey of Russian news media. With that said, this codebook provides ample documentation so that future researchers could, if they so choose, add additional news outlets of interest to make this sample more robust or suitable for their specific research interests.

Another potential pitfall of this database is that I scraped *all* articles available within each news outlet's archives. While I anticipate that many researchers will be primarily interested in the nature of political coverage, this database will inevitably include many articles about the arts, social issues, sports, or other themes that might be irrelevant for some research applications. To this end, the `Section` variable may be helpful. (One could, for instance, subset for "Politics" articles if they were interested in political coverage, or "Sports" articles if they were researching the prevalence of political themes in sports coverage.) With that said, some outlets do not provide section information, so this may require that researchers either accept some additional noise in their data, find another way to filter out irrelevant coverage, or omit outlets that do not provide section information.

Another major factor that I would urge users to take into consideration is the potential of purposely removed content. In many cases, the archives that I scraped go back several years, but I ultimately still conducted this scraping in late 2023; some content that was published in the past was likely taken down as a result of controversy, revision of outlet editorial policy, and so on. While this data can serve as an indicator of what was being published in past years, it should be noted that there is a chance that certain themes—say, favorable coverage of Yevgeny Prigozhin—might have been retroactively purged from outlets' archives as a result of changing political circumstances or other factors. Thus, analysis reaching further back should account for the potential of retroactive deletion.

One final consideration in thinking about the publication and use of this database is the question of intellectual property and copyright law. While I designed my methodology with best practices in mind, (I purposely slowed my scraping to avoid DoS issues, did not take articles from behind the few paywalls that I encountered, etc.) there is still some gray area about the

legality of republishing the text even of publicly available articles, even in a format as rudimentary as a csv file, as these articles are ultimately owned by the news outlets that published them. While the legality of various forms of web scraping is often murky, standing case law largely affirms that violations primarily occur as a result of scraping non-public information, republishing without reference to the original source, or because of site-specific user agreements that prohibit scraping. I only scraped publicly-available pages, provided links and outlet names for all articles documented, and signed no user agreements for any of the four outlets, but I would still likely pursue IRB or other legal counsel before publicizing this database to any outside researchers.

Appendix A. — Sitemap Configurations

The sitemaps provided below can be input into Web Scraper by using the "import sitemap" feature to achieve the same "raw" csv files that I wrangled for this database. These sitemaps essentially give Web Scraper a "lay of the land" for each outlet's web design, telling Web Scraper which html elements are headlines, when to press a "load more" button in the archives, and so on.

Single Home Page Site Map Example (Dozhd)

For some sites, a single start URL can be used to click through multiple archives:


```
{"_id":"Dozhd","startUrl":["https://tvrain.tv/archive/"],"selectors":[{"id":"More","parentSelectors":["_root","More"],"paginationType":"linkFromHref","type":"SelectorPagination","selector":"a.pagination__item:nth-of-type(n+2)"},{"id":"Article_Link","parentSelectors":["_root","More"],"type":"SelectorLink","selector":"a.chrono_list__item__info__name","multiple":true,"linkType":"linkFromHref"},{"id":"Headline","parentSelectors":["Article_Link"],"type":"SelectorText","selector":"h1","multiple":false,"regex":""},{"id":"Date","parentSelectors":["Article_Link"],"type":"SelectorText","selector":"span.document-head__date","multiple":false,"regex":""},{"id":"Views","parentSelectors":["Article_Link"],"type":"SelectorText","selector":".document-head__views
span","multiple":false,"regex":""},{"id":"Article_Text","parentSelectors":["Article_Link"],"type":"SelectorGroup","selector":"#article_content_text p","extractAttribute":""}]}
```

Multiple Home Page Example (Pravda)

For some sites, it is easier to parse through by month. See for instance, this map that scrapes through six months of Pravda's articles from 2021 using six different start URLs:


```
{"_id":"Pravda_2021","startUrl":["https://www.pravda.ru/archive/2021-01-[01-31]/","https://www.pravda.ru/archive/2021-02-[01-28]/","https://www.pravda.ru/archive/2021-03-[01-31]/","https://www.pravda.ru/archive/2021-04-[01-30]/","https://www.pravda.ru/archive/2021-05-[01-31]/","https://www.pravda.ru/archive/2021-06-[01-30]/"],"selectors":[{"id":"Article_Link","linkType":"linkFromHref","multiple":true,"parentSelectors":["_root"],"selec
```

tor":".news a, .six .column .title a, .video.article .title a","type":"SelectorLink"},{"id":"Headline","multiple":false,"parentSelectors":["Article_Link"],"regex":"","selector":"h1.title","type":"SelectorText"},{"extractAttribute":"","id":"Article_Text","parentSelectors":["Article_Link"],"selector":"h2, blockquote p, .full p, ol li, ul li, #default-list-item .full > p, #default-list-item p a, #default-list-item blockquote p","type":"SelectorGroup"},{"id":"Date","multiple":false,"parentSelectors":["Article_Link"],"regex":"","selector":"time[datetime='2019-01-01T21:01:00Z'], #default-list-item time","type":"SelectorText"}]}